

Natural Language
Assignment Project Exam Help
Pro tials

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro
Lizhen Q

What is Natural Language Processing?

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

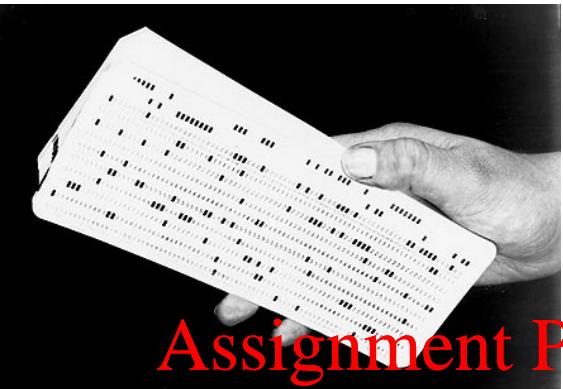
Solving engineering problems that need to analyze or generate natural language text.

Applications of NLP

- Speech recognition.
 - Question answering.
 - Machine Translation.
- Assignment Project Exam Help
- Spelling correcti
<https://eduassistpro.github.io/>
Add WeChat edu_assist_pro
 - Information extraction.
 - Summarization.
 - Dialogue systems.



History of Computational Linguistics



Assignment Project Exam Help

<https://eduassistpro.github.io/>



Add WeChat edu_assist_pro

1950s

Machine Translation

1960s

Syntax and QA

History of Computational Linguistics

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

1970s – 1980s

Semantics

1990s – present

Statistical NLP



Deep Learning

Google Assistant

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

https://www.youtube.com/watch?v=JvbHu_bVa_g

Google AutoML

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

<https://cloud.google.com/automl/>

Trump Speech

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

<https://filiph.github.io/markov/>

Obama's Speech

Good afternoon. God bless you.

The United States will step up to the cost of a new challenges of the American people that will share the fact that we created the problem. They were attacked and so that they have to say that all the task of the final days of war that I will not be able to get this done. The promise of the men and women who were still going to take out the fact that the to be able to protect our for the commitment to bo and women in uniform and the millions of our should be a strong stretches of the forces that w the American people and the leadership of our country who are on the Internet of American lives.

*Assignment Project Exam Help
https://eduassistpro.github.io/
Add WeChat edu_assist_pro
ke sure that they have
her to completely look
nd the fact is the men
law system that we
ncrease our spirit of*

Automatically Generate Text

- Machine generated talks.
 - <https://www.youtube.com/watch?v=-OodHtJ1saY>

Assignment Project Exam Help

- Machine gen
 - <http://www.c> <https://eduassistpro.github.io/> gi
 - <https://medium.com/@sami> Add WeChat edu_assist_pro machine-generated-political-speeches ea0#.56709ikih
 - <https://www.youtube.com/watch?v=EFHyzuqjaok>
 - <https://filiph.github.io/markov/>

Natural Language Grounding

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

What is the food?

Broccoli, rice, beans etc.

What color is the plate?

White

he green on the plate?
off

e broccoli?

late

Broccoli is on the white plate.

Red sauce is on the plate.

The plate is white.

Screenplay: <https://www.youtube.com/watch?v=LY7x2lhqjmc>

Levels of Language

Three years ago, Christina Wu was sitting at the Ivy Café in Canberra studying for her ANU exams. She tweeted about it, wondering if any of her friends were studying nearby.

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat **edu_assist_pro**

REQUEST: open the door.

STATEMENT: the door is open.

INFORMATION QUESTION: is the door open?

Levels of Language

- Phonetics and phonology – knowledge about linguistic sounds.
- Morphology – knowledge of the meaningful components of words.
- Syntax – knowledge of relationships between words <https://eduassistpro.github.io/>
- Semantics – knowledge of meaning
- Discourse – knowledge about units larger than a single utterance.
- Pragmatics – knowledge of the relationship of meaning to the goals and intentions of the speaker.

Assignment Project Exam Help

Add WeChat edu_assist_pro

Ambiguity of Natural Language

I made her duck.



Assignment Project Exam Help

- i. I cooked waterfowl.
- ii. I cooked water <https://eduassistpro.github.io/>
- iii. I created the duck she owns.
- iv. I caused her to quickly lower her body.
- v. I waved my magic wand and turned her into a waterfowl.



Overview of the NLP Lectures

- Introduction to natural language processing (NLP).
- Regular expressions, sentence splitting, tokenization, part-of-speech tagging.
[Assignment](#) [Project](#) [Exam](#) [Help](#)
- Language model <https://eduassistpro.github.io/>
[Add WeChat edu_assist_pro](#)
- Vector semantics.
- Parsing.
- Compositional semantics.

How to Find All Prices?

In January 2014, IBM announced plans to invest more than \$1.2bn (£735m) into its data centers and cloud storage business. It plans to build 15 new centers around the world, bringing the total number up to 40 during 2014.
[Assignment Project Exam Help](https://eduassistpro.github.io/)

In July 2014, the company announced it would invest \$3 billion over the following five years to make computer architecture resemble how the human brain thinks. A spokesman said: "Computer architecture has not altered since the 1940s. IBM is working to design a neural chip that mimics the human brain, with 10 billion neurons and 100 trillion synapses, but that uses just 1 kilowatt of power." <https://eduassistpro.github.io/>

Regular Expressions (RE)

- Definition: an regular expression is an algebraic notation for characterizing a set of strings.

pattern: ab^*c

Assignment Project Exam Help

abc ac adc
cd da aaa
bbb ccc ddd
acd aac Abc

<https://eduassistpro.github.io/>

abs aac adc
cd da aaa
bbb ccc ddd
acd aac Abc

Add WeChat [edu_assist_pro](#)

- RE Tools.
 - Perl, Java, Python etc.
 - grep, awk, vi, Emacs, Word etc.

RE Operators

- A single character: a, b, c ...
- A sequence of characters: country, fun ...
- []: disjunction of characters. E.g. [Cc]ountry .
- [-]: any one character in a range. E.g. [A-Z], [0-9] .
- *: zero or more <https://eduassistpro.github.io/> b *c
- +: one or more -9]+ .
- X{n}: X occurs exactly n times Add WeChat edu_assist_pro
- X{n,m}: X occurs at least n times but no more than m times.
- | : disjunction. E.g. parrot|bird .
- () : capturing group.

RE Demo

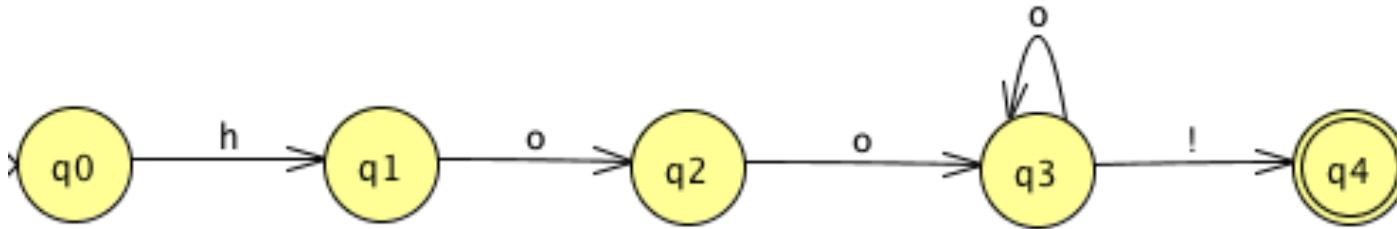
- Demo tool: <http://java-regex-tester.appspot.com/>

In January 2014, IBM announced plans to invest more than \$1.2bn (£735m) into its data centers and cloud storage business. It plans to build 15 new centers around the world, bringing the total number up to 40 during 2014.[82]

In July 2014, the co <https://eduassistpro.github.io/> ing \$3 billion over the following five years to create computer architecture to resemble how the human brain thinks. A spokesman said computer architecture had not altered since the 1940s. IBM says its goal is to design a neural chip that mimics the human brain, with 10 billion neurons and 100 trillion synapses, but that uses just 1 kilowatt of power.

- Search pattern?

Alternative Representations



- Regular expression Project Exam Help

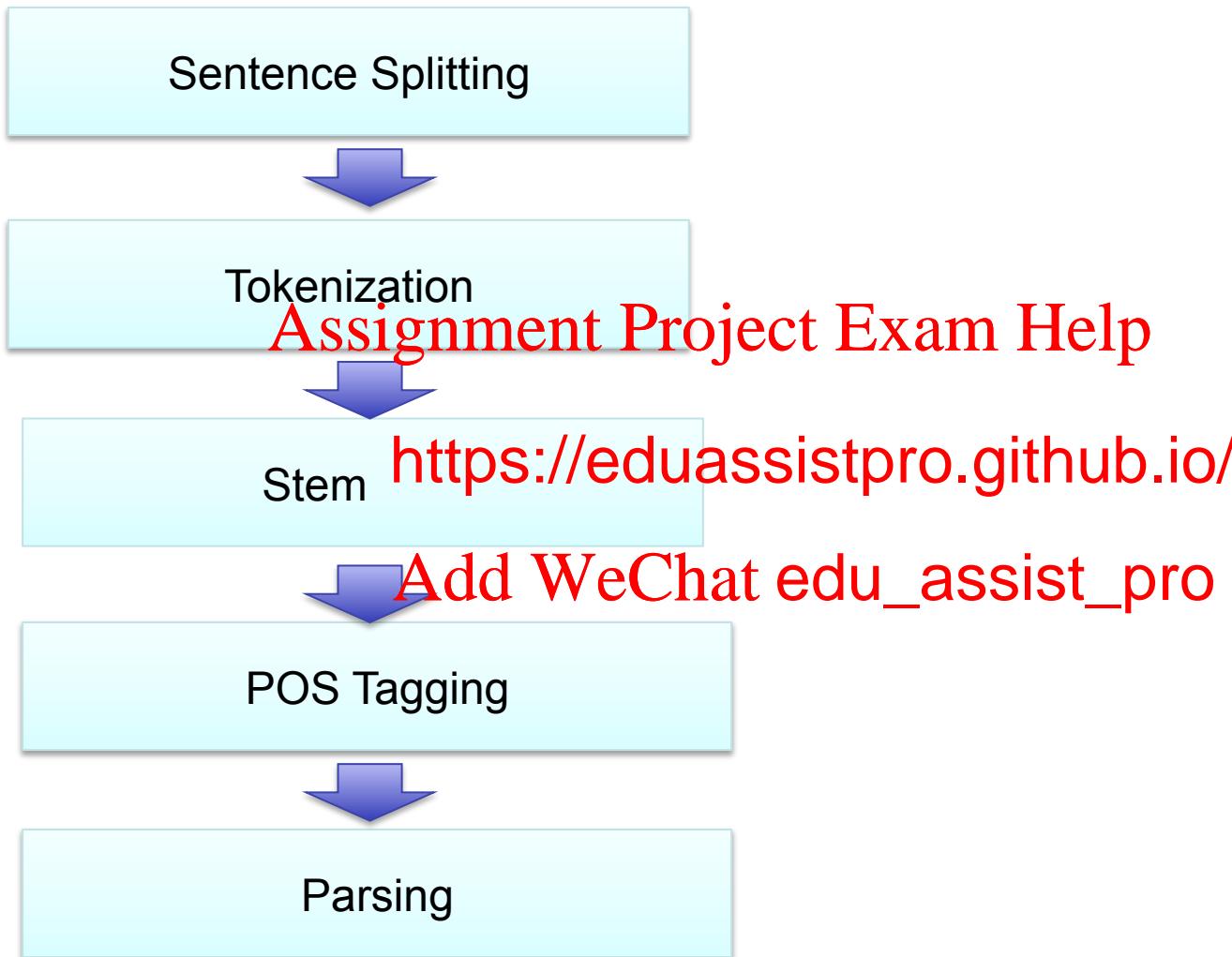
<https://eduassistpro.github.io/>

- State-transition

Add WeChat edu_assist_pro

	Input		
State	h	o	!
q0	q1	∅	∅
q1	∅	q2	∅
q2	∅	q3	∅
q3	∅	q3	q4
q4	∅	∅	∅

A Popular NLP Pipeline



Sentence Segmentation

- Problems with sentence splitting.
 - “You reminded me,” she remarked, “of your mother.”
- Sentence boundary detection algorithms.
 - Regular expressions.
 - Rule-based
 - Machine learning
- Tools.
 - OpenNLP.
 - Stanford CoreNLP.
 - NLTK.

Assignment Project Exam Help

<https://eduassistpro.github.io/>

Add WeChat edu_assist_pro

Tokenization

- Divide text into words, numbers, punctuations.
- English
 - Regular expressions?
E.g. [A-Za-z0-9]+|[\Punct]+
 - Problems.
 - Periods: E.
 - Clitic: E.g. i
 - Hyphenation.
 - *co-operate*
 - *most-visited* => *most visited*.
 - Tools.
 - OpenNLP.
 - Stanford CoreNLP.
 - NLTK.

Morphology

- Morphology is the study of the way words are built up from smaller meaning-bearing units.
 - Morpheme: minimal meaning-bearing unit in a language.

Assignment Project Exam Help
morphem <https://eduassistpro.github.io/>
stems **affixes** Add WeChat edu_assist_pro

fox, believe

prefixes
un-, sub-

suffixes
-s, -ed, -ly

infixes
hingi->humingi

circumfixes
üben -> geübt

Combining Morphemes to Create Words

- Inflection.
 - combination of a word stem with a grammatical morpheme.
 - same word class, e.g. *mouse* -> *mice*, *children's*, *walks*, *walked*, *ate*, *cut*.
- Derivation. [Assignment Project Exam Help](https://eduassistpro.github.io/)
 - combination of different word stems to create a new word with a different meaning.
 - different word stems can be combined to create a new word with a different meaning, e.g. *computation* -> [Add WeChat](https://eduassistpro.github.io/) [edu_assist_pro](https://eduassistpro.github.io/)
- Compounding.
 - combination of multiple word stems. E.g. *bedroom*.
- Cliticization.
 - combination of a word stem with a clitic. E.g. *I've* = *I* + *have*.
 - clitic: a morpheme that acts syntactically like a word, but is reduced in form and attached to another word.

Stemming (Review)

- A crude heuristic process that strips off suffixes.
- Algorithms.
 - Lookup algorithms.
 - Regular expressions.
 - Suffix-stripping
 - Porter stemmer <https://eduassistpro.github.io/>
<http://tartarus.org/martin/Porter>
 - Lovins stemmer [Add WeChat edu_assist_pro](http://snowball.tartarus.org/algorithms/lovins/stemmer.html)
<http://snowball.tartarus.org/algorithms/lovins/stemmer.html>
 - Lancaster stemmer.
<http://www.comp.lancs.ac.uk/computing/research/stemming/>
- Tools.
 - <http://snowball.tartarus.org>
 - NLTK

Porter stemmer.

- Lexicon free stemmer.
- Rewrite rules.
 - ATIONAL → ATE (e.g. *relational*, *relate*)
 - FUL → ε (e.g. *hopeful*, *hope*)
 - SSES → SS (e.g. *caresses*, *caress*)
- Errors of Com <https://eduassistpro.github.io/>
 - *Organization* → *organ*
 - *Policy* → *polic*
- Errors of Omission.
 - *urgency* → *urgenci* (not stemmed to *urgent*)
 - *European* → *European* (not stemmed to *Europe*)

English Parts of Speech

- **Noun** (person, place or thing)
 - Singular (NN): dog, fork
 - Plural (NNS): dogs, forks
 - Proper (NNP, NNPS): John, Springfields
 - Personal pronoun (PRP): I, you, he, she, it
 - Wh-pronoun (WP): who, what
- **Verb** (actions)
<https://eduassistpro.github.io/>
 - Base, infinitive
 - Past tense (VBD): ate
 - Gerund (VBG): eating
 - Past participle (VBN): eaten
 - Non 3rd person singular present tense (VBP): eat
 - 3rd person singular present tense: (VBZ): eats
 - Modal (MD): should, can
 - To (TO): to (to eat)

English Parts of Speech (cont.)

- **Adjective** (modify nouns)
 - Basic (JJ): red, tall
 - Comparative (JJR): redder, taller
 - Superlative (JJS): reddest, tallest
- **Adverb** (modify verbs)
 - Basic (RB): quickly
 - Comparative (R)
 - Superlative (R)<https://eduassistpro.github.io/>
- **Preposition** (IN): on, in, by, to
- **Determiner:** [Add WeChat edu_assist_pro](#)
 - Basic (DT) a, an, the
 - WH-determiner (WDT): which, that
- **Coordinating Conjunction** (CC): and, but, or...
- **Particle** (RP): off (took off), up (put up)

Closed vs. Open Class

- **Closed class** categories are composed of a small, fixed set of grammatical function words for a given language.
 - Pronouns, Prepositions, Modals, Determiners, Particles, Conjunctions.

Assignment Project Exam Help

<https://eduassistpro.github.io/>

- **Open class** categories have a large number of words and new ones are easily invented.
 - Nouns (Googler), Verbs (google), Adjectives (geeky), Adverb (chompingly)

Part of Speech Tagging

Assignment Project Exam Help

John saw the s https://eduassistpro.github.io/e it to the table.
NNP VBD DT B PRP IN DT NN

Add WeChat edu_assist_pro

English POS Tagsets

- Original Brown corpus used a large set of 87 POS tags.
- Most common in NLP today is the Penn Treebank set of 45 tags <https://eduassistpro.github.io/>
- Universal POS tags.
 - <http://universaldependencies.org/u/pos/>
 - <https://arxiv.org/pdf/1104.2086.pdf>

Ambiguity in POS Tagging

- “Like” can be a verb or a preposition.
 - I like/VBP candy.
 - Time flies like/IN an arrow.

Assignment Project Exam Help

- “Around” can be a verb, a preposition, or an adverb.
 - I bought it at the shop around/ticle, or/er.
 - I never got around/RP to gett Add WeChat [edu_assist_pro](https://eduassistpro.github.io/)
 - A new Prius costs around/RB \$25K.

Clues for POS Tagging

- Morphological clues.
 - -ment: government, establishment.
- Syntactic clues.
 - the *near* win
 - The end is v <https://eduassistpro.github.io/>
- POS in local context.
 - DET ? NN PUNC
 - DET NN BE ADV ?

POS tagger with LR: <http://nlp.stanford.edu/~manning/papers/emnlp2000.pdf>
Practice with NLTK: <http://www.nltk.org/book/ch05.html>

POS Tagger for Download

- Stanford POS tagger.
 - <http://nlp.stanford.edu/software/corenlp.shtml>
 - <http://nlp.stanford.edu/software/tagger.shtml>
- Open NLP.
[Assignment](#) [Project](#) [Exam](#) [Help](#)
- UIUC POS ta
 - <http://cogco> <https://eduassistpro.github.io/>
- NLTK.
[Add WeChat edu_assist_pro](#)
 - <http://text-processing.com/demo/tag>
- Twitter POS tagger.
 - <http://www.ark.cs.cmu.edu/TweetNLP/>
 - http://www.ark.cs.cmu.edu/TweetNLP/cluster_viewer.html

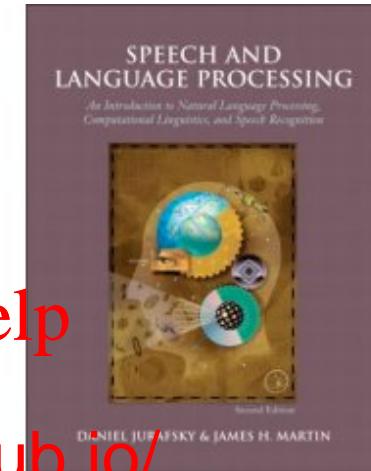
Textbooks

SPEECH and LANGUAGE PROCESSING

An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition

Assignment Project Exam Help
Daniel Jurafsky and James H. Martin

<https://eduassistpro.github.io/>
<https://web.stanford.edu>



Add WeChat edu_assist_pro

| Recommend to read at least 4.1

| <https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf>

| before the next class.